CAS ETH in Machine Learning in Finance and Insurance - Innovation project

Yield Crop Estimation Based on Remote Sensing Data

Project report

01st November 2024

Introduction

Problem description

Predicting crop yield is an important task from multiple perspectives. On the global and administrative level it's relevant for food security under the impact of growing population and climate change. It is expected that globally, food demand will increase between 2010 and 2050 by 35% to 56% while the population at risk of hunger will change from -91% to 30%[1]. On the local level accurate estimation of crop yield could help select the best season and plants to grow. There are multiple risk factors that affect the crop yield and that are difficult to manage on the individual level and hence create an opportunity for insurance and reinsurance companies to support and help diversify risk on the national and global level.

The primary aim of this project is to apply machine learning models to satellite images to predict corn crop yield in the USA. The secondary objective is to support the strategic goal of Blue Marble to build knowledge and vision models that would allow them to extend insurance coverage beyond strictly weather risks. As parametric insurance is the main business driver of this investigation, model performance for poor harvest years is of special interest.

Blue Marble is an Impact InsurTech with a mission to bring insurance to the underserved, backed by leading insurance organizations. They are present in Latin America, South and South-East Asia and Africa supporting local communities of farmers.

[1] van Dijk, M., Morley, T., Rau, M.L. *et al.* A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050. *Nat Food* 2, 494–501 (2021). https://doi.org/10.1038/s43016-021-00322-9

Literature discussion

Due to its importance crop yield prediction is a popular research topic and quite often featured in the top tier journals like Nature. Many methods have been used for this purpose. In recent years various machine learning methods have been applied to the problem. Systematic literature review from 2022[2] analysed and reviewed 44 articles while another one from 2020[3] another 80. According to both studies the popular architectures for the task are CNN and LSTM.

Due popularity of those two architectures and attractive dimensionality reduction property of histograms the decision has been made to attempt to reproduce results from paper by You, Jiaxuan, et al. "Deep gaussian process for crop yield prediction based on remote sensing data." Proceedings of the AAAI conference on artificial intelligence. Vol. 31. No. 1. 2017.

When performing literature review for this project two additional papers seemed particularly interesting due to application of transformers. Rad, Ryan. "Vision Transformer for Multispectral Satellite Imagery: Advancing Landcover Classification." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024 and Lin, Fudong, et al. "MMST-ViT: Climate Change-aware Crop Yield Prediction via Multi-Modal Spatial-Temporal Vision Transformer." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023. Both approaches could be potentially combined to improve prediction results.

[2] Muruganantham, Priyanga, Santoso Wibowo, Srimannarayana Grandhi, Nahidul Hoque Samrat, and Nahina Islam. 2022. "A Systematic Literature Review on Crop Yield Prediction with Deep Learning and Remote Sensing" Remote Sensing 14, no. 9: 1990. https://doi.org/10.3390/rs14091990

[3] *Thomas van Klompenburg, Ayalew Kassahun, Cagatay Catal, "Crop yield prediction using machine learning: A systematic literature review", Computers and Electronics in Agriculture, Volume 177, 2020, 105709, ISSN 0168-1699*

Modeling

High level description

We start by making an assumption that the location of a pixel in an image is not as important as its color. This allows us to reduce dimensions of the data significantly, group pixel values together and work with the counts of pixels i.e. histograms rather than full images themselves.

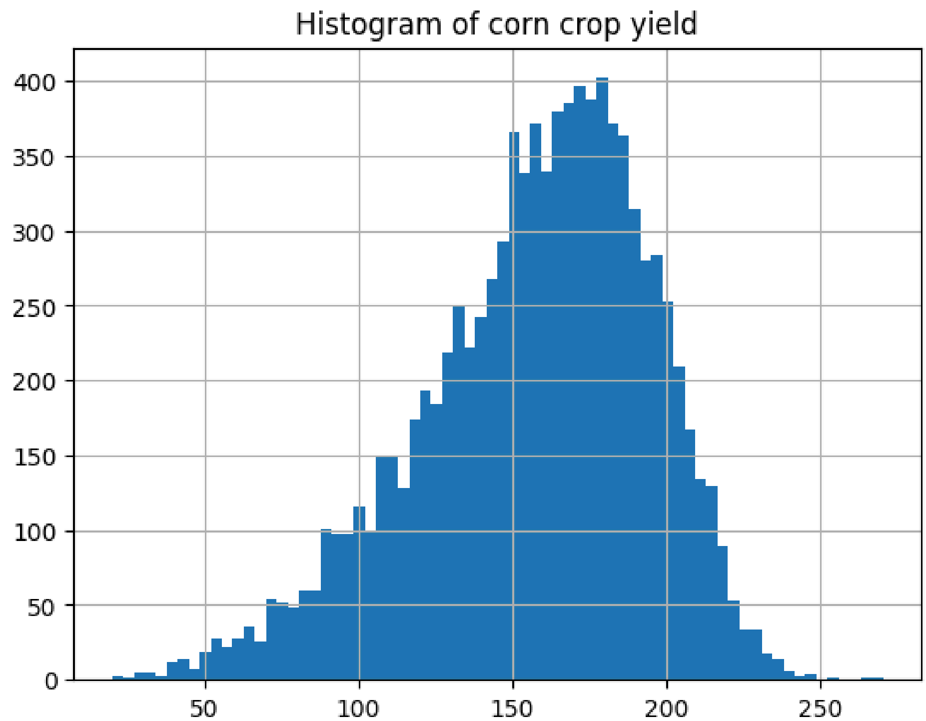
To make data collection possible within the timeframe of the project we make further simplifying assumptions that enough information will be captured at pixel resolution of 60x60 meters and that it's possible to reduce corn season lasting between May and September to three two month periods that can be represented by the median value of each pixel. It means that we look at 3 images per year, each pixel has a resolution of 60 by 60 meters and its value is a median value of all images that had less than 35% of cloudy pixels captured by Sentinel2 satellite within each two month period.

Data collection

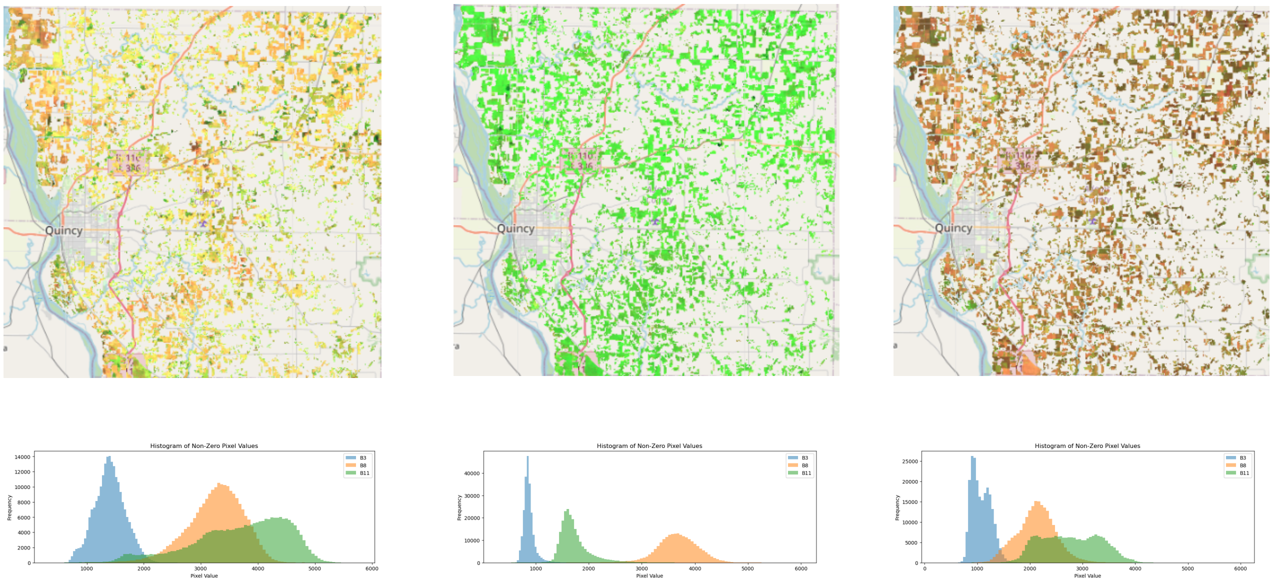
Labels were obtained from CropNet[4] dataset that was recently published on Hugging Face. The source of the labels is the United States Department of Agriculture and contains information on yearly corn yield measured in bushels per acre for the years 2016-2022. We choose the lowest level of aggregation possible and target predictions on the county level.

Interestingly, the distribution of labels is slightly skewed to the left with mean of 157, standard deviation of 38 and skewness of -0.58. Given that skewness is not significant we have decided not to apply any transformation to the data.

[4] CropNet/CropNet · Datasets at Hugging Face



We have collected 28661 Sentinel2 satellite images via google earth engine. To obtain a dataset that is in line with labels we cropped each image according to the county administrative borders based on 2018 census data and pixels representing corn crop according to the USDA National Agricultural Statistics Service according to the year. After cropping the images take roughly 80gb of storage space. The images were translated to the same amount of histograms. Subsequently all histograms from belonging to the specific county and year have been concatenated. In case one of the three histograms for the specific county and year combination were missing we used an empty histogram as a placeholder, otherwise the sample was discarded. In total this gives 9578 data points for regression. Histograms used for training are based on 9 channels: blue, green, red, red edge 1,2, 3 and 4 as well as near infrared and water vapor. Each histogram has 60 bins. These parameters were chosen as going much lower lead to poor model performance and going beyond did not much improvement and in extreme cases also lead to worse model performance.



[5]

Given that different seasons are clearly distinguishable the initial assumptions seem plausible. Perhaps even more so when we consider that each image has 13 channels in total.

Models, training and evaluation

We apply a typical test train split. We set aside 20% of data to use as a test dataset for model evaluation. We split the remainder further into training in validation datasets. Each holding respectively 64% and 16% of the initial data.

Three separate model architectures were implemented for comparison. Simple fully connected multi layer perceptron, LSTM and CNN. No simpler model has been proposed due to the high dimensionality of the problem.

General observations related to model architectures applied to this task. Based on loss function only there is no clear favorite among the three approaches.

MLP has proven to be very flexible and can easily work with a penalized loss function that was used with the intention of improving the fit in the left tail of the distribution. Indeed that allows for slightly better results for predicting bad years but comes with higher variance on the unseen data. LSTM architecture has proven to be difficult to work with. Most changes in the model architecture beyond adding an attention layer leads to

---

[5] Adams county Illinois, images representing the year 2017. From the left median of months May and June, July and August and September and October and their corresponding histograms. Visualised using green, short wave infrared and near infrared channels.

the vanishing gradients and constant predictions. Having said that, once the model finds the good path beyond the average the model offers the smoothest learning path. CNN had the advantage of being quite flexible without convergence issues. It has decent tail performance and lower variance in predictions.

Interestingly up to the moment of introducing state code in the training dataset , CNN model was performing marginally better than the other two. With the extra covariate however LSTM took the lead. It generalizes better and performs better in low yield scenarios. The improved model performance with additional covariate indicates that there are other important factors, potentially linked to the geography that are not covered by satellite images.

For brevity only LSTM architecture will be presented in detail. All experiments performed during this project are available for inspection at weights and biases[6] portal. Furthermore github repository[7] contains the whole codebase. Access to both is public.

LSTM model architecture

Model layers:

- 7 stacked LSTM layers with 200 units each, default activation functions
- Additive attention layer with 200 attention units
- Fully connected linear output layer

This LSTM model implementation has approximately 8mln parameters.

Training parameters:

- Learning rate: 0.00013
- 300 epochs with callback monitoring validation loss

Model performance on test data

Model	RMSE	MAE	Corr
LSTM	24	19	0.777
CNN	26.6	19	0.759
MLP	32.64	22	0.75

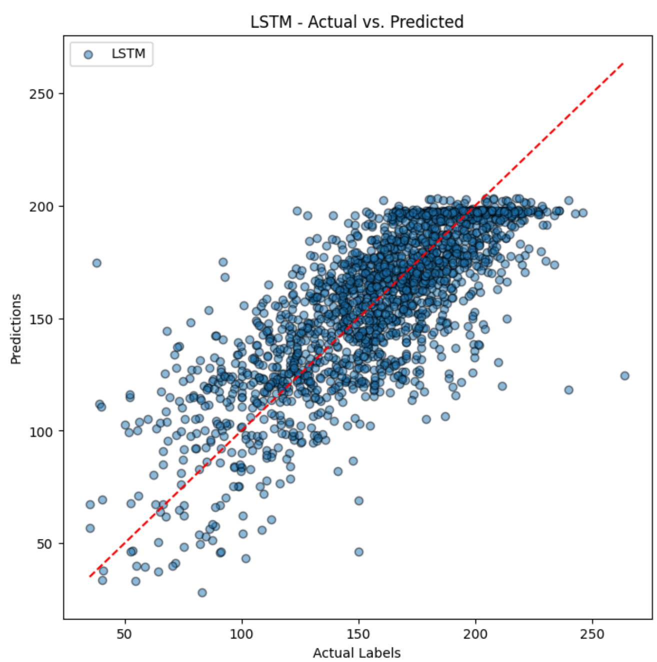
---

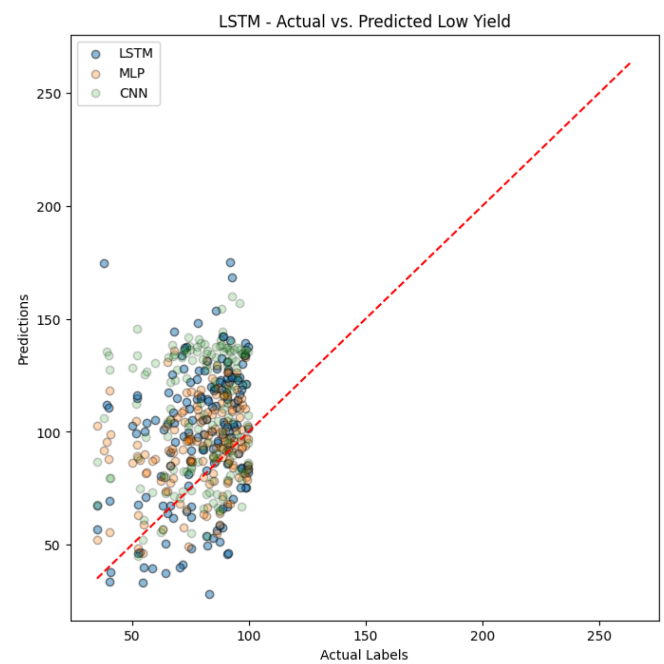
6 https://wandb.ai/t-skorkowski/blue-marble?nw=nwusertskorkowski

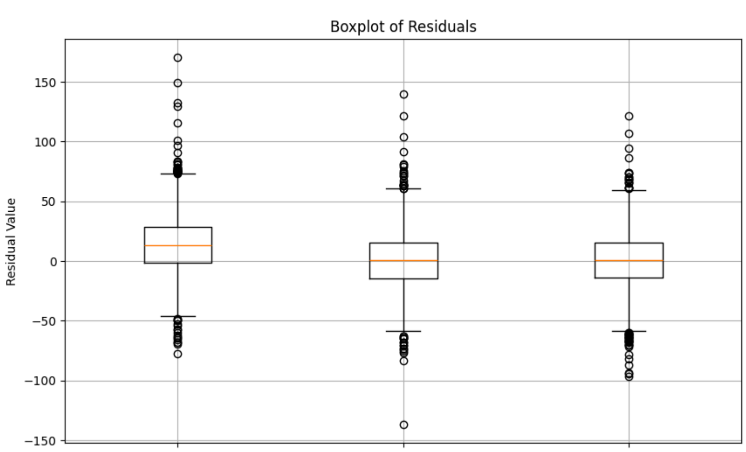
7 https://github.com/tskorkowski/crop_yield_prediction_CAS

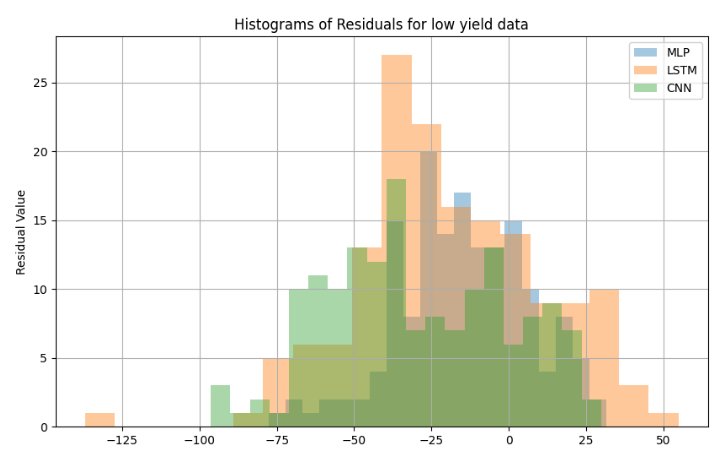
If we take 100 bushels per acre as a low crop threshold then 8.4% which is equivalent to 158 samples that fall into that category. Mean squared error for observations below this threshold is the lowest for MLP and equal to 7.45, followed by LSTM and 10.4 and finishing with CNN and RMSE of 12.

Looking at both residuals and model fit we observe that even though residuals are centered around the zero, all models tend to significantly overestimate low crop yields and underestimate good years. From a practical perspective that means that the models studied are not good enough yet to build an insurance product. Afterall with such a model the benefit payment would rarely be made to those who need it.









When compared to the reference study our models still perform significantly worse. Among other things this could be due to difference in channels from MODIS satellites or due to difference in temporal data. Nevertheless, the obtained results further confirm the value of satellite images as a data source.

Next steps

The idea of being able to predict crop yield seems too attractive to give up. There are multiple options on how to take this work forward. One way would be to capture higher resolution data. Due to time constraints this was impossible for this project but resolution can easily be increased to 10x10m meters or 20x20m, depending on the channel.

Another natural extension to the current work would be to apply CNN directly to the images instead of histograms.

Another attractive idea would be to replicate the two prospective papers mentioned in the literature discussion. This approach not only uses more recent model architecture but takes advantage of multiple data modalities as well.